Signal Processing Tool for Emotion Recognition

Manish Kumar Saini, J.S.Saini & Ravinder
Electrical Engineering Department
Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Sonepat, Haryana, India
Email ID: ravinder.7987@gmail.com

***Abstract* - In the course of realization of modern day robots, which not only perform tasks, but also behaves like human beings during their interaction with the natural environment, it is essential for us to impart knowledge of the underlying emotions in the spoken utterances of human beings to the robots, enabling them to be consistent, whole, complete and perfect. To this end, it is essential for them too to understand and identify the human emotions. For this reason, stress is laid now-a-days on the study of emotional content of the speech and accordingly speech emotion recognition engines have been proposed. This paper is a survey of the main aspects of speech emotion recognition, namely, features extractions and types of features commonly used, selection of most informed features from the original dataset of the features, and classification of the features according to different classifying techniques based on relative information regarding commonly used database for the speech emotion recognition.**

***Index terms* – S.E.R (speech emotion recognition), utterance, elicit.**

I. Introduction

The technique of speech emotion recognition refers to the process of recognizing the emotional state of the human beings, using speech signals only. Various kinds of information regarding speech signal such as acoustics, prosodic and sometimes linguistic, are used either separately or in combination with each other to carry out the task of speech emotion recognition. Specific features are extracted from the above mentioned types of information, regarding speech signals that are useful in detecting and recognizing emotions of the speaker. Some applications of speech emotion recognition are such as recording the frustration of the customer in ticket reservations systems, so that the system can change its response accordingly [1]. Speech emotion recognition systems have also proved to be very helpful in call centres [2], where these systems are employed to recognize the emotional state of the customer so that the system or, for that matter, the caller could adapt. A very popular application of speech emotion recognition is in aircrafts, where, the advanced pilot control system periodically monitors' the emotional state of the pilot [3].

Emotion recognition through speech is a three step process, where first of all feature extraction. Second step is feature selection. Finally, last step of the process is classification of utterance into one of the predefined emotional class.

This paper is divided into 9 sections. Section 2 provides scientific as well as linguistic meaning of an utterance. Section 3 gives detailed explanation about emotions and its representation in emotional space. Section 4 provides detailed information regarding some of very popular databases used for speech emotion recognition purpose. Section five deals with features used for the emotion classification. Section 6 provides brief preview of feature selection algorithms. Finally, section 7 provides information regarding different classifiers.

II. Utterance

An utterance is a natural unit of speech bounded by breaths or pauses of the speaker. Linguists sometimes refer to utterance as simply a unit of speech under study. The act of uttering or vocal act of expression is also termed as utterance. We use the term utterance to refer to complete communicative units, which may consist of single word, phrase or clause. In the process of speech emotion recognition, the features are normally extracted by either dividing an utterance into frame which consists of voiced or unvoiced intervals or for the complete utterance [4], [5].

III. Emotion

Although there is little consensus as to what defines emotion, yet generally emotion refers to conscious states that are experienced, which can be characterized primarily by biological changes, physiological expressions and mental states. For use in the field of research, researchers define emotion as a two or three dimensional space. The space of emotions of consist dimensions of activation and valence. Thus, any emotion can be plotted as a point or small area in the space of emotion. Energy spent or energy associated with a particular emotion refers to activation. For example, emotions of anger and happiness are associated with large amount of energy, while emotions of sadness or neutrality are associated with low amount of energy. Second dimension of emotional space is valence.

Valence provides the information regarding the positivity or negativity associated with the particular emotion. For example, sadness is associated with a lot of negativity while happiness is associated with a lot of positivity.

Arousal of sympathetic nervous system is directly associated with the emotions. As with the emotions of joy, anger and fear, a large amount of energy is related to arousal of concerned nervous system, which in turn results in decreased salivation, high blood pressure, increased heart rate, etc. This in turn produces speech which is characterized as loud and fast, with high energy content at higher frequencies. On the other hand, with emotions like sadness, low blood pressure and decreased heart rate are associated with increased salivation, with low energy content at higher frequencies [6].

DOI: 03.LSCS.2013.3.30

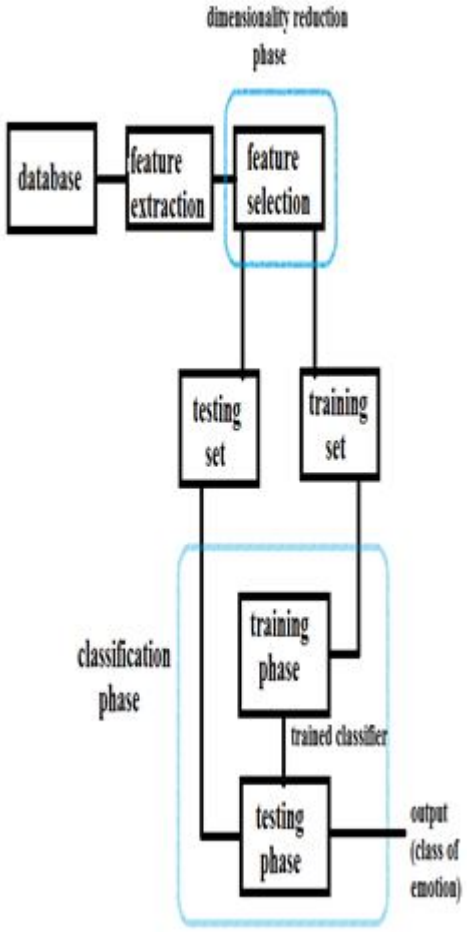

Figure 1: Block diagram for speech emotion recognition

IV. Database

Percentage of successful recognition of a particular emotion in an utterance by the S.E.R engine depends upon the naturalness of the database used to perform the task. Commonly used databases for the emotion recognition are mostly acted ones or the elicited ones. Databases created in studios or with the help of professionals are mainly used [5]. These kinds of databases are termed as artificial databases. There are mainly two kinds of artificial databases available. First one is simulated or acted speech database, in which professional actors are used to utter the emotional speech sentences. Normally, several professional actors are used to utter the same sentence in different emotions. As in real life situations, the content of the spoken words sometimes doesn't matter, instead human beings recognize the meaning of the dialogue in the context of underlying emotion in that dialogue. Second type of database used is the elicited speech database, non professional actors when try to elicit the sentences with emotions, the database created by these recordings is termed as elicited speech database. For example, when people try to elicit dialogues spoken by the film stars in the films,There are several well known emotional databases available, but only some of them are freely accessible. Several problems exist in the emotional speech databases available such as:

1. Different databases are composed of utterances of some particular type of emotions only. Therefore, a universal database still does not exist. So we have to either use same database for training and testing or we have to use utterances of all the databases with common emotions only. Universality in training and testing is lost.
2. The quality of database recordings of all the databases is not the same, which could affect the performance of the S.E.R engine.
3. Descriptions regarding linguistic information and phonetic information are not provided in some databases, which pose problems when hybrid mechanisms are used for emotion recognitions [9].

Some commonly used databases for speech emotion recognition are *Danish emotional speech* [7], *Berlin emotional speech database* [8], *eEnterface* [9], *SUSAS* [10], *KISMET* [22].

V. Features For Speech Emotion Recognition

A. Statistical Features

The features used are maximum and minimum values and respective relative positions, Range (max-min), Arithmetic mean and quadratic mean, Zero crossing and mean crossing rate, Arithmetic mean of peaks, Centriod of contour, Number of peaks and mean distance between peaks, Linear regression coefficients and corresponding approximate error, Quadratic regression error, Standard deviation, Variance, skewness and kurtosis, Arithmatic, Quadratic and geometric means of absolute and non zero values

B. Low Level Descriptors

Energy - (log energy), Pitch - fundamental frequency (F0), Zero crossing rate, Probability of voicing, Spectral energy – F bands.

C. MFCC

Mel-frequency cepstral coefficients, 2nd and 3rd order coefficients of MFCC, 2nd and 3rd order derivative of MFCC.

D. Average of Mean Log Spectrum.

E. Perceptual Linear Predictive Coefficients.

F. Nonlinear Teager energy operator Based Features (TE0)

TEO-FM-VAR (TEO decomposed FM variation),
TEO-AUTO-ENV (autocorrelation envelope area),
TEO-CB-AUTO-ENV (critical band based TEO-AUTO-ENV).

Features for the speech emotional recognition can be termed as information regarding the emotional content of the speech hidden in the speech signal. Studies have shown that the quality and the type of the features used for the emotional recognition from the speech signal have great impact on the efficiency of the SER engine used [12], [13].

Selection of the feature set to be used for speech emotion recognition also depends upon the type of classifier used for the classification task in the emotion recognition through speech [14], [15].

During the earlier phase of research in the field of emotion recognition through speech, development of new and

DOI: 03.LSCS.2013.3.30

advanced classification systems takes the edge over other areas of development in this field, but in the recent phase, idea to select the optimal feature set, which increases the overall efficiency of successful recognition of the emotions in the speech signal [12].

Features extracted on the utterance level are normally termed as global features. Global features have shown better results than local features [16], [17], [18]. Global features are very few in numbers, so analysing them is faster and more efficient. Local features are extracted on frame level. Some new features are also proposed by researchers such as excitation source features [31], class level spectral features [14], glottal features [32], New Harmony features [34], and modulation spectral features [33].

VI. Feature Selection

Feature selection phase acquires a lot of attention of researchers now-a-days in the field of speech emotion recognition. There are two types of feature selection algorithms available namely filter type algorithm and wrapper type algorithm [21], [20], [19].

Several strategies are formulated in terms of feature selection frameworks to obtain optimum feature subset for the classification purpose. In these frameworks, strategies are employed like one-vs.-rest and one-vs.-one strategy [12].

VII. Classification

The last step of speech emotion recognition is classification. It involves classifying the raw data in the form of utterance or frame of the utterance into particular class of emotion on the basis of features extracted from the data. There are several types of conventional classifiers available for research in this field namely Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Artificial Neural Network (ANN) etc.

A. Multiple or Hierarchical Classifier Systems [30], [36], [37].
B. Hidden Markov model [23].
C. *Gaussian Mixture [24]*.
D. Support Vector [25].
E. Artificial Neural Networks [26], [35].

Some other types of classifiers are also proposed by some researchers like 3DEC hierarchical classifier in which highly confused classes are separated first from rest of the classes and the procedure is repeated until all the classes are separated. Another proposed strategy is optimum-path forest classification in which training set is taken as a graph, whose nodes are samples and link weights are determined by the distance between the feature vectors of their corresponding nodes [29].

VIII. Conclusion

In this survey paper, the key aspects of speech emotion recognition engine have been studied. Types of databases available, with important characteristics of each database are provided. Commonly used features for the speech emotion recognition and some new features proposed in recent studies with their better classification results are also mentioned. Types of feature selection algorithms are discussed. A sample of the hierarchical classifier strategy is also provided for future work.

References

[1] Ang, J, Dhillon, R. Krupski, A. Shriberg and E. Stolcke, "Prosody based automatic detection of annoyance and frustration in human computer dialogue," *International Conference on Spoken Language Processing* (ICSLP '02), vol. 3, pp. 2037-2040, 2002.

[2] Petrushin V.A, "Emotion in speech recognition and application to call centres," *Artificial Neural Network in Engineering* (ANNIE '99), vol. 1, pp. 7-10, 1999.

[3] J. Hansen and D. Cairns, "Icarus: source generator based real-time recognition of speech in noisy stressful and Lombard effect environment," *Speech Communication* 16(4), pp. 391-422, 1995.

[4] R. Fernandez, *A computational model for the automatic recognition of affect in speech*, Ph. D thesis, M.I.T, February, 2004.

[5] C. Williams and K. Stevens, "Vocal correlates of emotional states" *Speech Evaluation in Psychiatry*, Grune and Stratton," pp. 1238-1250, 1981.

[6] J.Chan,"The generation of affect in synthesized speech," *J. Am. Voice input/output* soc. 8, pp. 1-19, 1990.

[7] I.S. Engbert and A.**V** .Hansen, "Documentation of Danish emotional speech database DES," Technical report, *Center for Person Communication*, Aalborg university, Denmark, 2007.

[8] F. Burkhardt, A. Paeschke, M. Rolfes and W. Sendlmier, and B. Weiss, "A database of German emotional speech," *Proceedings EUROSPEECH*, pp. 1517-1520, 2005.

[9] O. Martin I. Kostia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," *Proceedings IEEE Workshop Multimedia Database Management*, 2006.

[10] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," *EUROSPEECH*, vol. 4, pp. 1743-1746, 1997

[11] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth and G. Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, 2010.

[12] Haltis Altu and, Gokhan Polat, "Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection," *Expert Systems with Applications*, vol. 36, 2009.

[13] Lijiang Chen, Xia Mao and Yuli Xue, Lee lung Cheng, "Speech emotion recognition: features and classifications," *Digital Signal Processing*, pp. 1051-2004, 2012.

[14] D. Bitouk, Ragini verma and Ani nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, pp. 613-625, 2010.

[15] O. Kwon, K. Chan, J. Hao and T. Lee, "Emotion recognition by speech signal,"*EUROSPEECH*, Geneva, pp. 125-128, 2003.

[16] M.T. Shami and M.S. Kamel, "Segment-based approach to the recognition of emotions in speech," *IEEE International Conference on Multimedia and Expo, ICME 2005*, 4pp, 2005.

[17] H.Hu,M.-X.Xu,W.Wu, "Fusion of global statistical and segmental spectral features for speech emotion recognition," in*: International Speech Communication Association 8th Annual Conference of the International Speech Communication*

DOI: 03.LSCS.2013.3.30

Association, Interspeech 2007, vol.2, 2007, pp.1013–1016.

[18] D. Ververidi and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm," *IEEE International Conference on Multimedia and Expo, ICME* 2005, pp. 1500–1503, 2005.

[19] R. Caruana, and D. Freitag, "Greedy attribute selection," *11th International Conference on Machine Learning*, pp. 153-172, 1994.

[20] M. Dash, K. Choi and P. Scheurmann, and H. Liu, "Feature selection for clustering a filtering solution," *2nd International Conference on Data Mining*, pp. 115-122, 2002.

[21] Huan liu, and Lei Yu, "Towards integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, 2005.

[22] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot directed speech," *Autonomous Robots* 2, pp. 83–104, 2002.

[23] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE Transactions on Acoustic, Speech, and Signal Processing Magazine. 3* (1), pp. 4–16, 1986.

[24] N. Vlassis and A. Likas, "A greedy expectation-maximization (EM) algorithm for Gaussian mixture learning," *Neural Process, Lett.* 15, pp. 77–87, 2002.

[25] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995

[26] R. Duda, P. Hart, D. Stork, *Pattern Recognition*, John Wiley and Sons, 2001.

[27] Chung-Hsien, and Wei-Bin Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, 2011.

[28] L. Rabiner and B. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

[29] Hassan and R. I. Damper, "Classification of emotional speech using 3-DEC hierarchical classifier," *Speech Communication*, vol. 54, pp. 903-916, 2012.

[30] Enrique M. Albornoz, Deigo H. Milone, Hugo L. Rufiner, " Spoken emotion recognition using hierarchical classifiers," *Computer Speech and Language*, vol. 25, pp. 556-570, 2011.

[31] Shashidhar G. Koolgudi, Swati Devliyal, Bhavna Chawla, Anurag Bartwal, K. Sreenivasan rao, "Recognition of emotions from speech using excitation source features," *Procedia Engineering*, vol. 38, pp. 3409-3417, 2012.

[32] Alexander I. Iliev, Michael S. Scordilis, Joao P. Papa and Alexandre X. Falcao, "Spoken emotion recognition through optimum path forest classification using glottal features," *Computer Speech and Language*, vol. 24, pp. 445-460, 2010.

[33] Siqing Wu, Tiago H. Falk, wai-Yip Chan, "Automaic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, pp. 768-785, 2011.

[34] B. Yang and M. Lugger, "Emotion recognition from speech using New Harmony features," *Signal Processing*, vol. 90. Pp. 1415-1423, 2010.

[35] Joy Nicholson, Kazuhiko Takahashi and Ryohie Nakatsu, "Emotion recognition in speech using neural network," *IEEE*, 1999.

[36] Liquin Fu, Xia mao and Lijiang Chen, "Speaker independent emotion recognition based on SVM/HMMs fusion systems," *IEEE*, 2008.

[37] Mingyu You, Chun Chen, Jiajun Bu, Jia Liu and Jianhua Tao, "A hierarchical framework for speech emotion recognition," *IEEE*, 2006, Canada.

DOI: 03.LSCS.2013.3.30